Information Integration and the Semantic Web*

Tim Finin and **Anupam Joshi**
University of Maryland, Baltimore County
Baltimore MD
(finin,joshi)@umbc.edu

Li Ding
Stanford University
Stanford CA
ding@ksl.stanford.edu

The Semantic Web vision is to develop a Web of data. Independent agents (people and programs) publish information on the Web using RDF, a simple, common representation language with a well defined semantics. The documents represent ontologies (conceptual models) or data. Ontologies are machine interpretable definitions of classes and properties identified with unique URIs. Data documents use terms defined in ontologies to materialize RDF graphs making assertions about Web resources, again identified by unique URIs. This approach provides a good foundation for publishing data that is designed to be integrated.

Unpacking the phrase *Semantic Web* immediately produces its two constituent concepts: it is (i) a semantic framework to represent the meaning of data that is (ii) designed for use on the Web. Most current research, basic and applied, has focused on the first of these and largely ignored the second. An obvious lesson from the last ten years of Web-based developments is we must not underestimate the impact of the (still emerging) Web on technology and society.

Current research includes many projects on all aspects of RDF and OWL as knowledge representation languages – complexity, scalability, completeness, efficient reasoning algorithms, integration with databases, rule extensions, etc. as well as work on systems and tools for ontology engineering, visualization, manual markup, etc. Application papers typically center on using RDF to express the knowledge and data for particular problem domains, such as workflow models, healthcare records, or policies. For the most part, current work touches little on issues that stem for the (initial) intended use of Semantic Web languages for publishing and using ontologies and data on the World Wide Web.

Much practical work has been done, of course, to develop Web appropriate standards for the Semantic Web and harmonizing them with existing Web standards and practices. Many applications and test beds have also focused on core Web paradigms, such as semantically enhanced Web services and policy-driven negotiation for Web resource access. Our claim is that we need more research on modeling and understanding how Semantic Web concepts and technology is and can be used on the Web. In this respect, we stand on the shoulders of those who call for *"Creating a Science of the Web"* (Berners-Lee *et al.* 2006).

*Partial support was provided by NSF awards ITR-IIS-0326460 and ITR-IDM-0219649.

Swoogle. The Swoogle Semantic Web search system[1] (Ding *et al.* 2004; 2005b) discovers, analyzes and indexes Semantic Web documents on the Web. At this writing, it has processed more than 1.7 Million documents comprising more than 310 Million RDF triples. Just as modern search engines and their services have given people access to much of the world's knowledge and data, Swoogle aims to support Semantic Web developers, software agents and programs in finding RDF data and schema-level knowledge on the Web. We have used Swoogle as a component in a number of information integration tasks and are working on several others. We briefly describe some of these use cases below.

Supporting Semantic Web developers. The first and most common use of Swoogle is to support developers and researchers. Swoogle helps developers in finding ontologies and individual RDF terms (i.e., classes and properties) for study and reuse, in finding data to illustrate how these ontologies have been used, and in finding RDF data of interest. For example, once can use Swoogle to find what terms have been used to represent a person's email address, how much each such property has been used and by what documents. This helps developers choose ontologies and terms likely to be understood by others and promotes the emergence of consensus ontologies.

Helping scientists publish and find data. Sharing data is extremely important in the natural sciences and experimental engineering disciplines. The Semantic Web offers new ways for scientists and engineers to publish and find both data and associated ontologies (Finin & Sachs 2004). In (Aleman-Meza *et al.* 2006) we explored problems in integrating disparate data about authors, papers, institutions, and collaborations in support of a system that discovered and ranked evidence for potential *conflicts of interest* in the context of matching reviewers to papers submitted to a conference or journal. We have been working with a group of biologists to develop systems to allow them to share ecoinformatics models and data (Parr *et al.* 2006; Sachs *et al.* 2006). Their data can easily be exported and published as RDF from applications, spreadsheets and conventional databases.

Helping Semantic Web researchers. Swoogle also helps researchers who study how the Semantic Web is being used.

[1] http://swoogle.umbc.edu/

For example, researchers can find what properties have been used with a particular class, like *foaf:Person*, including properties that violate constraints associated with the *foaf* (Friend of a Friend) ontology. Swoogle's database lets language designers see what features of RDF and OWL are used (and misused!) in practice. Swoogle also maintains an archive of all versions of documents in it's index, allowing one to model how ontologies and data change and evolve on the Semantic Web (Wang, Parsia, & Hendler 2006). For a widely used semantic system, it likely that the semantics will drift, to some degree, by social forces. Swoogle can help track and monitor use and inform standards bodies considering revisions. Swoogle's global catalog of terms opens up new opportunities to rethink some fundamental design issues, like whether ontology documents are required, as opposed to collections of (loosely coupled) class and property definitions.

Discovering ontology mappings. Swoogle can also be used to support ontology mapping. Large ontologies like Cyc and WordNet are unlikely to have complete mappings to other ontologies. Swoogle can be used to assemble partial ontology mappings from multiple sources by collecting assertions specifying mappings expressed using OWL primitives (e.g., *owl:sameClass*) or terms from special ontology mapping ontologies. Swoogle can also be used to compile instance data for terms in different ontologies that can then be used to induce mapping relationships, as in (Pan *et al.* 2005).

Learning trust relationships. Swoogle can be used to provide evidence for trust relationships based on who is using what ontologies and what data. When integrated with other metadata and Semantic Web data, interesting relationships can perhaps be derived.

Finding proofs. We've used Swoogle to maintain and access a special collection of reasoning proofs marked up in PML, *proof markup language*. This allows applications to find proofs in support of a particular fact. Using the notion of an *RDF molecule* (Ding *et al.* 2005a), these proofs can be "strengthened" by finding additional sources on the Semantic Web that provide support for premises.

Discovering facts. Swoogle can be used to collect data matching certain patterns, e.g., find all RDF triples asserting facts about a foaf:Person instance with foaf:lastName equal to "Finin".

Feeding queries. SPARQL (Prud'hommeaux & Seaborne 2006) is a being developed by the W3C as the first standard query language for RDF. Like SQL, it has a *FROM clause* that identifies one or more RDF documents over whose combined graphs the query is run. We have implemented a experimental system (Sachs *et al.* 2006) allowing a user to compose a SPQRQL query without specifying the data sources. Swoogle is used to find RDF documents that contain data relevant to the query and that match additional constraints posed by the user (e.g., trustworthiness). The query can be run against the collected corpus or it can be materialized and saved as dataset for later use.

Conclusion. Web search engines and their services have provided essential infrastructure enabling people to find and integrate information expressed in natural languages on the Web. Software agents can benefit from similar search engines and services designed to discover, analyze, index, and retrieve information encoded in Semantic Web languages like RDF and OWL. The Swoogle Semantic Web search system has discovered and processed more than 1.7M RDF documents published on the Web. We have explored how its services can be used for information discovery and integration in a number of applications.

References

Aleman-Meza, B.; Nagarajan, M.; Ramakrishnan, C.; Sheth, A.; Arpinar, B.; Ding, L.; Kolari, P.; Joshi, A.; and Finin, T. 2006. Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection. In *WWW 2006, Edinburgh, Scotland*.

Berners-Lee, T.; Hall, W.; Hendler, J.; Shadbolt, N.; and Weitzner, D. J. 2006. Creating a science of the web. *Science* 313(5788):769–771.

Ding, L., and Finin, T. 2006. Characterizing the semantic web on the web. In *Proceedings of the 5th International Semantic Web Conference*.

Ding, L.; Finin, T.; Joshi, A.; Pan, R.; Cost, R. S.; Peng, Y.; Reddivari, P.; Doshi, V. C.; and Sachs, J. 2004. Swoogle: A Search and Metadata Engine for the Semantic Web. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*. ACM Press.

Ding, L.; Finin, T.; Joshi, A.; Peng, Y.; da Silva, P. P.; and McGuinness, D. L. 2005a. Tracking RDF Graph Provenance using RDF Molecules. In *Proceedings of the 4th International Semantic Web Conference*.

Ding, L.; Finin, T.; Joshi, A.; Peng, Y.; Pan, R.; and Reddivari, P. 2005b. Search on the Semantic Web. *IEEE Computer* 10(38):62–69.

Ding, L.; Pan, R.; Finin, T.; Joshi, A.; Peng, Y.; and Kolari, P. 2005c. Finding and ranking knowledge on the semantic web. In *Proceedings of the 4th International Semantic Web Conference*.

Ding, L.; Zhou, L.; Finin, T.; and Joshi, A. 2005d. How the semantic web is being used:an analysis of FOAF. In *Proceedings of the 38th International Conference on System Sciences*.

Finin, T., and Sachs, J. 2004. Will the Semantic Web Change Science? *Science Next Wave*.

Pan, R.; Ding, Z.; Yu, Y.; and Peng, Y. 2005. A Bayesian Network Approach to Ontology Mapping. In *Proceedings of the Fourth International Semantic Web Conference*.

Parr, C.; Parafiynyk, A.; Sachs, J.; Ding, L.; Dornbush, S.; Finin, T.; Wang, T.; and Hollender, A. 2006. Integrating Ecoinformatics Resources on the Semantic Web. In *Proceedings of the 15th International World Wide Web Conference*. Poster paper.

Prud'hommeaux, E., and Seaborne, A. 2006. SPARQL query language for RDF. http://www.w3.org/TR/rdf-sparql-query/. (accessed 5 September 2006).

Sachs, J.; Parr, C.; Parafiynyk, A.; Pan, R.; Han, L.; Ding, L.; Finin, T.; Hollender, A.; and Wang, T. 2006. Using the Semantic Web to Support Ecoinformatics. In *Proceedings of the AAAI Fall Symposium on the Semantic Web for Collaborative Knowledge Acquisition*. American Association for Artificial Intelligence.

Wang, T.; Parsia, B.; and Hendler, J. 2006. A survey of the web ontology landscape. In *Proceedings of the 5th international semantic web conference*.